

## Data Education for Children

Welcome to our data science education project! Today's learning outcomes are listed below:

1. Knowing what is Data Science
2. Understanding what the data set looks like
3. Knowing 4 basic data visualization techniques
4. Designing data visualizations regarding the real data
5. Knowing how to read an infographic

Please assign each of you a specific role as follows:

*Manager:* keeps track of time and makes sure everyone contributes appropriately.

*Recorder:* records all answers and questions, so team members and the facilitator have accurate notes.

You will first read materials, then answer the questions. Please write down your answers **on the answer sheet** if you see . **Do not** write anything on the teaching material. Stop continuing if you see . If you are stuck by any questions, just ask the facilitator for help.

## -Exploration-

Let's see three examples regarding Data Science in real life.

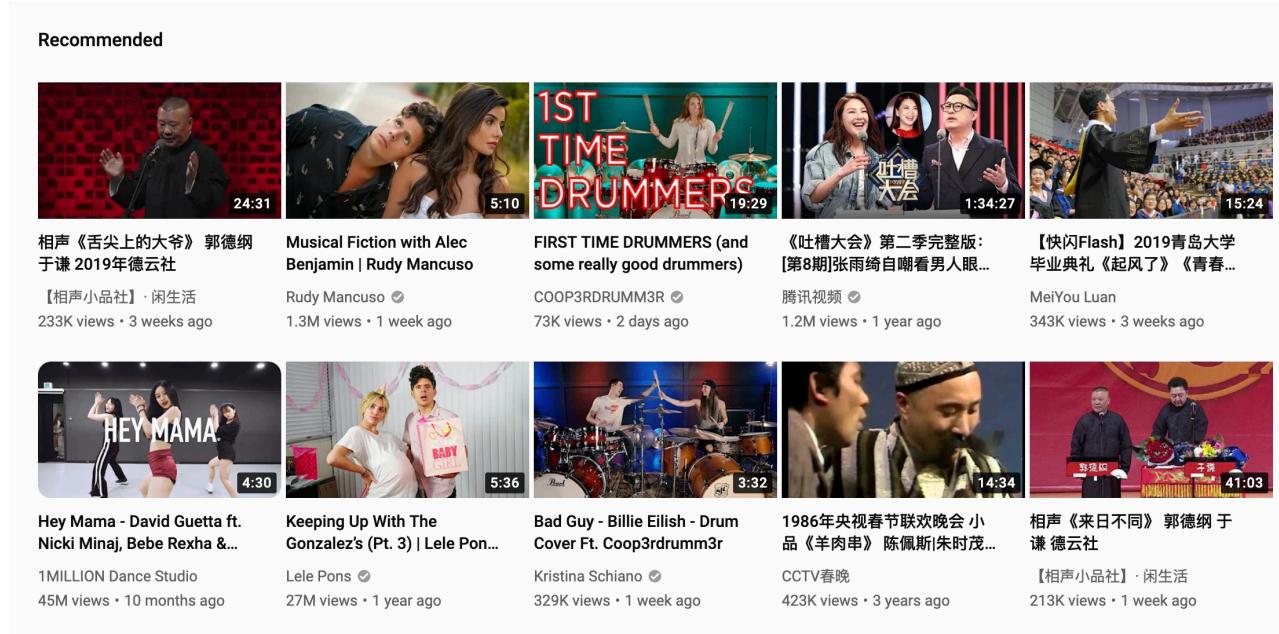


Figure 1: the recommendation system of Youtube

(3 min) Question 1: Why these recommendations are not like yours? What data do you think is needed for the recommendation system? 🤔



Figure 2: the weather forecast of Edinburgh.

(3 min) Question 2: How do you think the scientists predict the weather? ↗

**Figure 3: the friends suggestion of Facebook.**

(3 min) Question 3: How do you think Facebook predict the people you may know? ↗

(4 min) Question 4: Use your word to conclude how does Data Science work. ↗

Now, let's have a look at today's first dataset.

	quality_score	Species	Owner	Country of Origin	Farm Name	Lot Number	Mill
0	90.58	Arabica	metad plc	Ethiopia	METAD PLC		METAD PLC
1	89.92	Arabica	metad plc	Ethiopia	METAD PLC		METAD PLC
2	89.75	Arabica	Grounds for Health Admin	Guatemala	San Marcos Barrancas "San Cristobal Cuch		
3	89	Arabica	Yidnekachew Dabessa	Ethiopia	Yidnekachew Dabessa Coffee Plantation		Wolensu
4	88.83	Arabica	metad plc	Ethiopia	METAD PLC		METAD PLC
5	88.83	Arabica	Ji-Ae Ahn	Brazil			
6	88.75	Arabica	Hugo Valdivia	Peru	n/a		HVC
7	88.67	Arabica	Ethiopia Commodity Exchange	Ethiopia	Aolme		C.P.W.E
8	88.42	Arabica	Ethiopia Commodity Exchange	Ethiopia	Aolme		C.P.W.E
9	88.25	Arabica	Diamond Enterprise Plc	Ethiopia	Tulla Coffee Farm		Tulla Coffee
10	88.08	Arabica	Mohammed Lalo	Ethiopia	Fahem Coffee Plantation		
11	87.92	Arabica	CQI Q Coffee Sample Representative	United States	El filo		
12	87.92	Arabica	CQI Q Coffee Sample Representative	United States	Los Cedros		
13	87.92	Arabica	Grounds for Health Admin	United States (Hawaii)	Arianna Farms		
14	87.83	Arabica	Ethiopia Commodity Exchange	Ethiopia	Aolme		C.P.W.E
15	87.58	Arabica	CQI Q Coffee Sample Representative	United States	El Águila		

**Table 1: The dataset regarding Arabica coffee beans**

We got the information regarding 1312 Arabica and 28 Robusta coffee beans in **two independent datasets**. The 1312 Arabica coffee beans are from 35 countries and regions. The 28 Robusta coffee beans are from 5 countries. This is a screenshot of Arabica's dataset. The dataset contains many aspects of coffee beans. **We want to compare their quality by countries of origin, based on quality scores.** Thus, what we need is **species, country of origin, and the quality point.**

We need to do data preprocessing first - bin the useless data and transform the raw data in a useful and efficient form. We keep species, country of origin, and quality score.

	Species	Country.of.Origin	Total.Cup.Points
1	Arabica	Ethiopia	90.58
2	Arabica	Ethiopia	89.92
3	Arabica	Guatemala	89.75
4	Arabica	Ethiopia	89
5	Arabica	Ethiopia	88.83
6	Arabica	Brazil	88.83
7	Arabica	Peru	88.75
8	Arabica	Ethiopia	88.67
9	Arabica	Ethiopia	88.42
10	Arabica	Ethiopia	88.25
11	Arabica	Ethiopia	88.08
12	Arabica	United States	87.92
13	Arabica	United States	87.92
14	Arabica	United States (Hawaii)	87.92
15	Arabica	Ethiopia	87.83

**Table 2: the preprocessed dataset.**

(2 min) Question 5: Compare the qualities of the first 5 coffee beans regarding the *Total.Cup.Points* in the table below. ↗

	Species	Country.of.Origin	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Points	Total.Cup.Points
1	Arabica	Guatemala	8.42	8.5	8.42	8.42	8.33	8.42	10	10	10	9.25	89.75
2	Arabica	Ethiopia	8.75	8.67	8.5	8.58	8.42	8.42	10	10	10	8.58	89.92
3	Arabica	Ethiopia	8.25	8.5	8.25	8.5	8.42	8.33	10	10	10	8.58	88.83
4	Arabica	Ethiopia	8.67	8.83	8.67	8.75	8.5	8.42	10	10	10	8.75	90.58
5	Arabica	Ethiopia	8.17	8.58	8.42	8.42	8.5	8.25	10	10	10	8.67	89

**Table 3: the first 5 coffee beans**

(3 min) Question 6: Is there a faster way other than just looking at the figures - what can we transform the figures to? Hint: recall Figure 2. 🤝

(3 min) Question 7: We got 1312 Arabica coffee beans from 35 countries and regions (Table 2 only shows 15 of them). What statistical term do we use to compare the quality of their coffee beans by country and region? 🤝 Review status with the facilitator before continuing. 🧑

(3 min) Question 8: The average scores of the coffee beans from Colombia, India, China, Uganda, United States, Ecuador, Ethiopia, Japan are provided below. Complete the bar chart. 🤝 Review status with the facilitator before continuing. 🧑

	Country of origin	Average Score
1	Colombia	83.11
2	India	76.83
3	China	82.93
4	Uganda	84.05
5	United States	85.98
6	Ecuador	83.83
7	Ethiopia	84.54
8	Japan	84.54

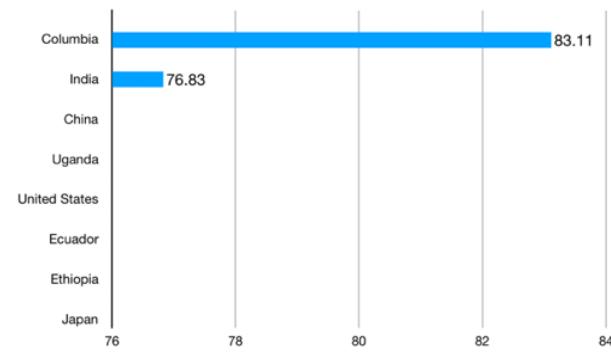
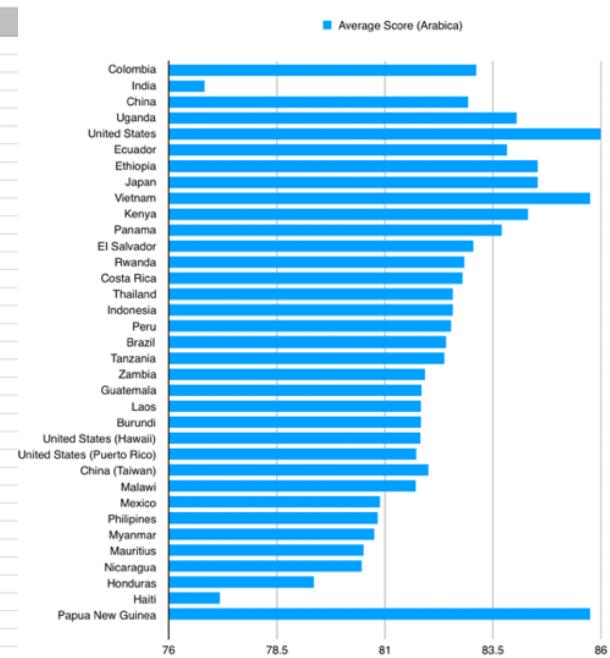


Table 4

Figure 4

Next, we calculate the average scores of 35 countries and regions and get the entire table and visualisation.

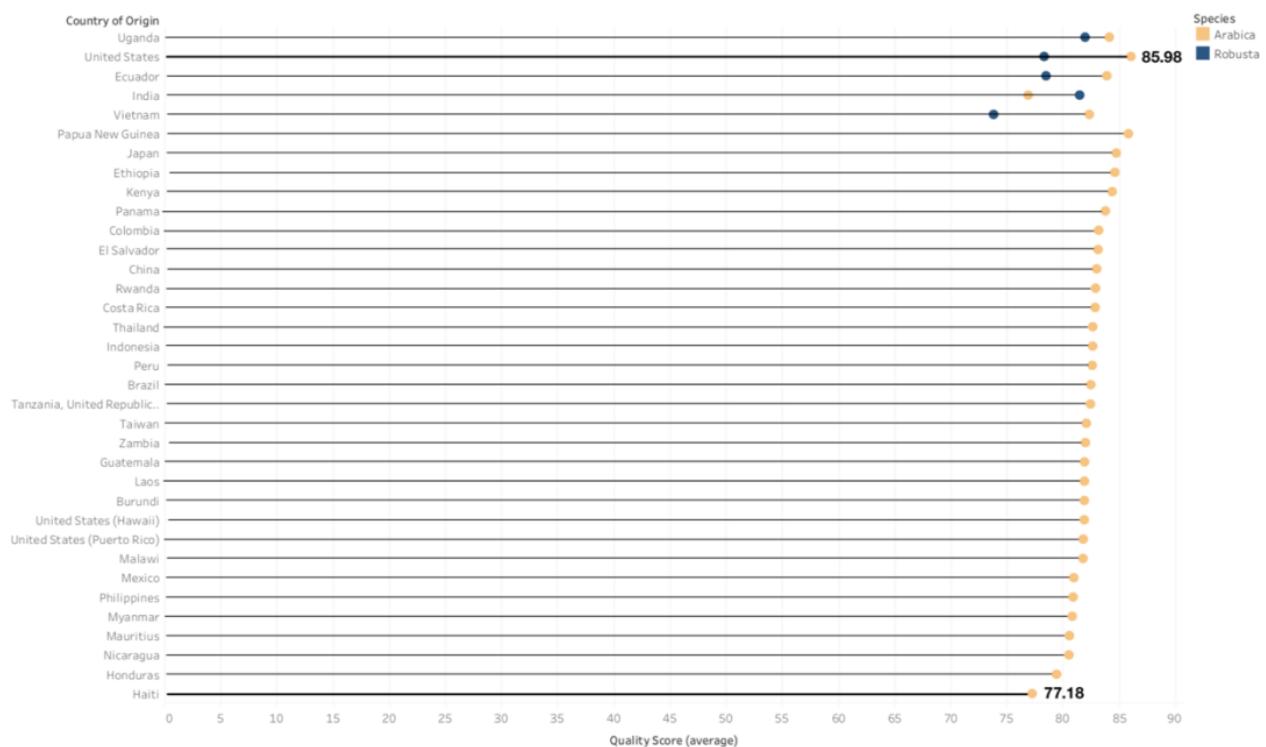
Country of origin	Average Score (Arabica)
1 Colombia	83.11
2 India	76.83
3 China	82.93
4 Uganda	84.05
5 United States	85.98
6 Ecuador	83.83
7 Ethiopia	84.54
8 Japan	84.54
9 Vietnam	85.75
10 Kenya	84.31
11 Panama	83.71
12 El Salvador	83.05
13 Rwanda	82.83
14 Costa Rica	82.79
15 Thailand	82.57
16 Indonesia	82.57
17 Peru	82.53
18 Brazil	82.41
19 Tanzania	82.37
20 Zambia	81.92
21 Guatemala	81.85
22 Laos	81.83
23 Burundi	81.83
24 United States (Hawaii)	81.82
25 United States (Puerto Rico)	81.73
26 China (Taiwan)	82.00
27 Malawi	81.71
28 Mexico	80.89
29 Philippines	80.83
30 Myanmar	80.75
31 Mauritius	80.50
32 Nicaragua	80.46
33 Honduras	79.36
34 Haiti	77.18
35 Papua New Guinea	85.75



**Table 5****Figure 5**

Do you remember there are **two independent datasets**? The visualisation above is just for Arabica coffee beans. We still need to complement the quality scores of 28 Robusta coffee beans from 5 countries to our data visualisation (this step is to ensure the integrity of the full cycle of data visualisation. You do not need to complement anything).

Average Quality Scores of Coffee Beans from Different Countries (Regions)

**Figure 6: average quality scores of coffee beans from different countries and regions.**

You have understood how to do data preprocessing and data visualisation so far. Next section is to learn three widely used visualisation techniques. Review status with the facilitator before continuing.

## -Concept Invention-

(5 min) Question 9: Read these two pie charts below and find out what is in common in the two datasets.

Pie Chart #1.

English dialect	Proportion
American	70.7%
British	15.9%
Canadian	4.9%
Australian	4.8%
Other	3.7%

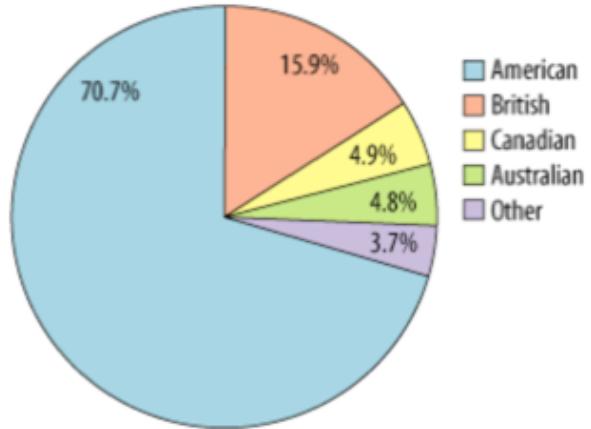


Table 6

Figure 7

The table and chart show the relative numbers of native English speakers in the major English-speaking countries of the world.

Pie Chart #2.

End Use	proportion
Shower	16.8%
Toilet	26.7%
Leaks	13.7%
Tap	15.7%
Washing Machine	21.7%
Other	5.3%

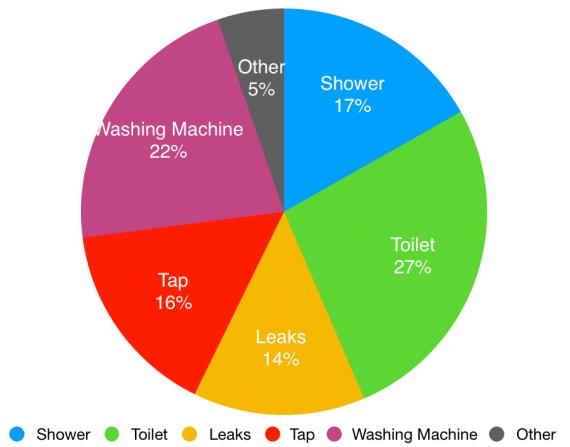


Table 7

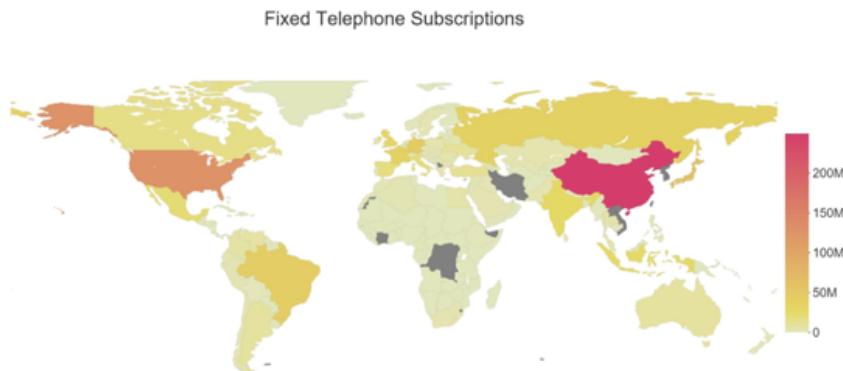
Figure 8

The table and chart show residential end uses of water in 1999. Now please answer question 9. ↗

(5 min) Question 10: Read these two map charts below and find out what is in common in the two datasets.

Map Chart #1.

	Country	Phone Lines
1	Afghanistan	101931.0
2	Albania	235734.0
3	Algeria	3098787.0
4	American Samoa	9900.0
5	Andorra	38237.0
6	Angola	281327.0
7	Antigua and Barbuda	19915.0
8	Argentina	9602605.0
9	Armenia	572772.0
10	Aruba	35000.0
11	Australia	9190000.0
12	Austria	3254700.0
13	Azerbaijan	1795448.0
14	Bahamas	125658.0
15	Bahrain	284684.0
16	Bangladesh	974181.0



Source: World Telecommunication/ICT Indicators Database

Table 8

Figure 9

The table and chart show the distribution of fixed telephone subscription in the world.

Map Chart #2.

Name	population
London	7556900
Birmingham	984333
Liverpool	864122
Nottingham	729977
Sheffield	685368
Bristol	617280
Glasgow	591620
Leicester	508916
Edinburgh	464990
Leeds	455123
Cardiff	447287
Manchester	395515
Stoke-on-Trent	372775

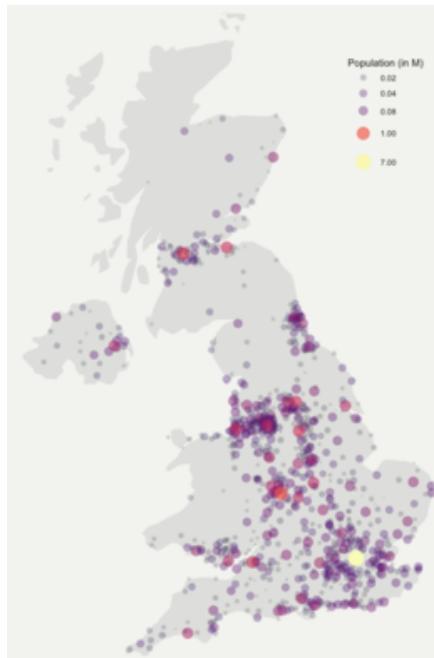


Table 9

Figure 10

The table and chart show the 1000 biggest cities and towns in the UK by population in 2019. Now please answer question 10.

(5 min) Question 11: Read these two map charts below and find out what is in common in the two datasets.

Line Chart #1.

Time	Temperature
00:00, Sun 30 Jun	14
01:00, Sun 30 Jun	15
02:00, Sun 30 Jun	15
03:00, Sun 30 Jun	15
04:00, Sun 30 Jun	14
05:00, Sun 30 Jun	14
06:00, Sun 30 Jun	15
07:00, Sun 30 Jun	15
08:00, Sun 30 Jun	16
09:00, Sun 30 Jun	16
10:00, Sun 30 Jun	18
11:00, Sun 30 Jun	16
12:00, Sun 30 Jun	17
13:00, Sun 30 Jun	18

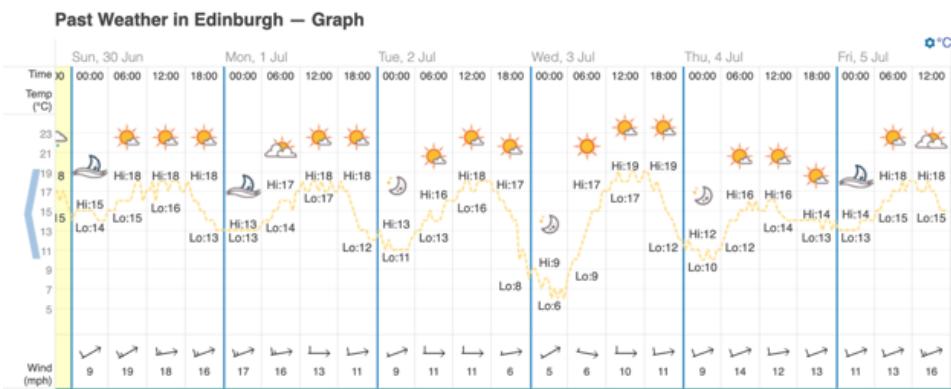


Table 10

Figure 11

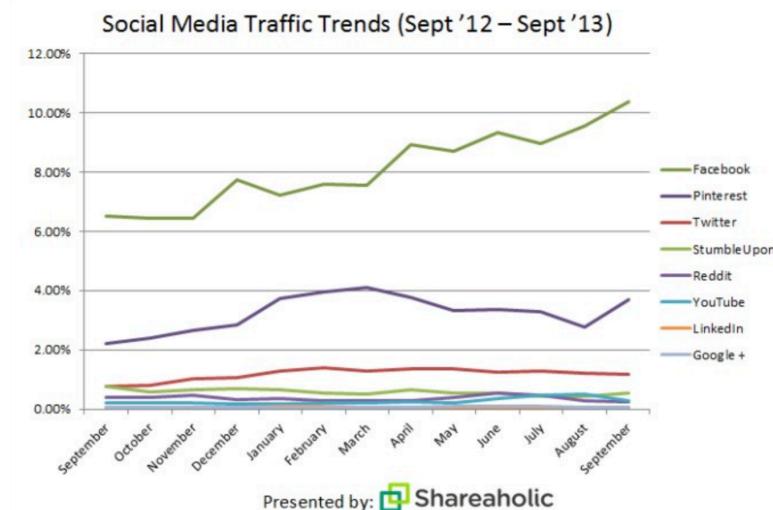
The table and chart show the weather of Edinburgh from 30/06 to 05/07.

Line Chart #2.

Social Media Traffic Referrals (September 2012 – September 2013)															
Source	Share of visits September	Share of visits October	Share of visits November	Share of visits December	Share of visits January	Share of visits February	Share of visits March	Share of visits April	Share of visits May	Share of visits June	Share of visits July	Share of visits August	Share of visits September	13-month average	Growth from Sept'12-Sept'13
Facebook	6.53%	6.45%	6.45%	7.76%	7.24%	7.61%	7.56%	8.95%	8.70%	9.34%	8.95%	9.58%	10.37%	8.11%	58.81%
Pinterest	2.21%	2.38%	2.67%	2.84%	3.73%	3.94%	4.09%	3.76%	3.33%	3.38%	3.29%	2.77%	3.68%	3.24%	66.52%
Twitter	0.76%	0.80%	1.01%	1.08%	1.30%	1.39%	1.28%	1.35%	1.35%	1.25%	1.30%	1.20%	1.17%	1.17%	54.12%
StumbleUpon	0.77%	0.58%	0.64%	0.69%	0.65%	0.55%	0.51%	0.65%	0.56%	0.53%	0.44%	0.43%	0.56%	0.58%	-27.47%
Reddit	0.40%	0.40%	0.46%	0.33%	0.37%	0.30%	0.27%	0.28%	0.40%	0.55%	0.45%	0.27%	0.26%	0.36%	35.16%
YouTube	0.19%	0.21%	0.21%	0.18%	0.18%	0.19%	0.22%	0.23%	0.22%	0.35%	0.46%	0.49%	0.29%	0.26%	52.86%
LinkedIn	0.05%	0.06%	0.06%	0.06%	0.08%	0.08%	0.07%	0.07%	0.08%	0.07%	0.06%	0.07%	0.07%	0.07%	34.51%
Google +	0.04%	0.05%	0.06%	0.05%	0.05%	0.06%	0.06%	0.07%	0.08%	0.08%	0.08%	0.05%	0.04%	0.06%	6.97%

Presented by: Shareaholic

Table 11



## Figure 12

The table and chart show social media traffic trends (Sept'12 - Sept'13). Now please answer question 11. ↗

(4 min) Question 12: Based on the commonalities in each of the visualisation techniques you just find out, conclude in what situations do we use pie chart, map chart, and line chart? ↗ Review status with the facilitator before continuing. 🤝

(5 min) Question 13: Based on the answer of question 12, what data visualization other than bar chart can we use to compare the quality of coffee beans by different countries and regions? Why? Sketch the visualisation to verify your assumption. ↗ Review status with the facilitator before continuing. 🤝

We have learned bar chart, pie chart, line chart, and map chart so far. In next section we will experience the full cycle of data visualisation - data collection, data entry, data preprocessing, data visualisation, using the real data you create.

## **-Application-**

Let's finish a small task with jelly beans.

(2 min) Question 14: Open jelly beans and taste some of them.

(12 min) Question 15: In the remaining chocolate, select one or more features (ask the facilitator if you are not sure about what 'feature' is). Then generate your dataset (in the table form) and express it in a visualization way that you think is appropriate. 

	Feature 1	Feature 2	Feature 3	...
1				
2				
3				
4				
5				
6				
...				

**Table 12: the template of the table.**

(4 min) Question 16: do you find something interesting in your visualisation? Review status with the facilitator before continuing. 

(8 min) Question 17: Read the infographic below following the three steps: 1) read the topic; (2) find the panel that interests you; (3) read the panel. If you are stuck, try to find some explanation near your interested panel. Is there something interesting in this infographic? 

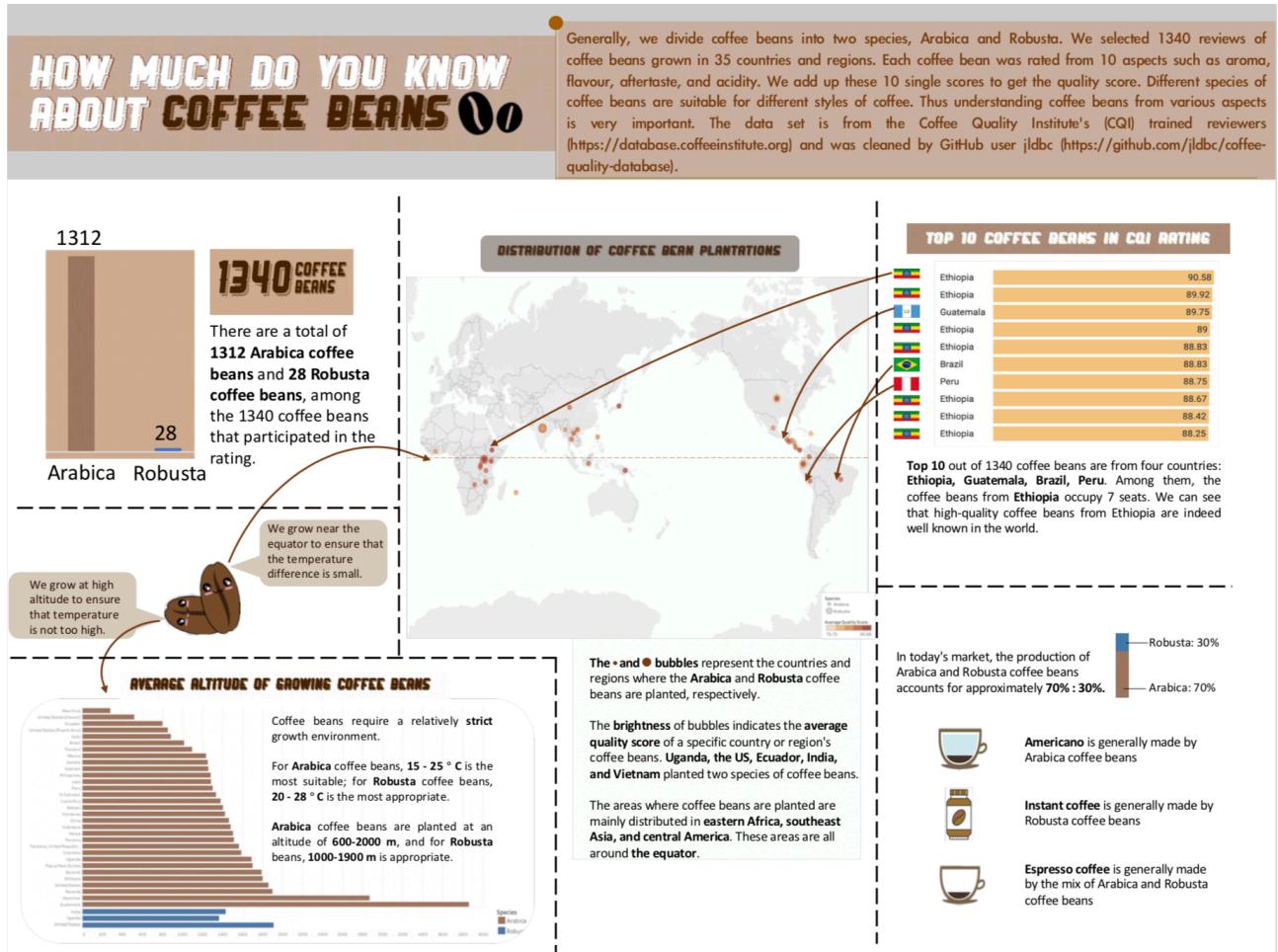


Figure 13: infographic about coffee beans

So far, we have learned what is data science, and we have experienced the full cycle of data visualisation. Also, we understand how to read an infographic now. With the concepts we learned today, we can generate data visualisation using real world data, and read all the hot news related to data science!